Special Section: Observational Studies

Ways of Learning: Observational Studies Versus Experiments

TERRY L. SHAFFER,[1] *United States Geological Survey, Northern Prairie Wildlife Research Center, 8711 37th Street SE, Jamestown, ND 58401, USA*

DOUGLAS H. JOHNSON, *United States Geological Survey, Northern Prairie Wildlife Research Center, 204 Hodson Hall, 1980 Folwell Avenue, Saint Paul, MN 55108, USA*

ABSTRACT Manipulative experimentation that features random assignment of treatments, replication, and controls is an effective way to determine causal relationships. Wildlife ecologists, however, often must take a more passive approach to investigating causality. Their observational studies lack one or more of the 3 cornerstones of experimentation: controls, randomization, and replication. Although an observational study can be analyzed similarly to an experiment, one is less certain that the presumed treatment actually caused the observed response. Because the investigator does not actively manipulate the system, the chance that something other than the treatment caused the observed results is increased. We reviewed observational studies and contrasted them with experiments and, to a lesser extent, sample surveys. We identified features that distinguish each method of learning and illustrate or discuss some complications that may arise when analyzing results of observational studies. Findings from observational studies are prone to bias. Investigators can reduce the chance of reaching erroneous conclusions by formulating a priori hypotheses that can be pursued multiple ways and by evaluating the sensitivity of study conclusions to biases of various magnitudes. In the end, however, professional judgment that considers all available evidence is necessary to render a decision regarding causality based on observational studies. (JOURNAL OF WILDLIFE MANAGEMENT 72(1):4–13; 2008)

DOI: 10.2193/2007-293

KEY WORDS association, bias, causation, designed experiment, learning, matching, monitoring, observational study, randomization, sample survey.

The word *management* in the title of this journal suggests that we care about causality. Wildlife managers perform some management action and would like wildlife to respond in a law-like, predictable manner. But causality is a more challenging concept in our field than in, for example, the physical sciences, where models of the behavior of atoms, planets, and other inanimate objects are applicable over a wide range of conditions (Barnard 1982), and there are few controlling factors. Outside of some of the physical sciences, however, notions of causality reduce to those of probability. Causation then means that an action "tends to make the consequence more likely, not absolutely certain" (Pearl 2000:1). This is so in wildlife ecology because of the multitude of factors influencing a system. As an example, liberalizing hunting regulations is expected to increase harvest by hunters. In any single instance, however, liberalization may not cause a greater harvest because other influences impinge, such as the number of animals in the population, weather conditions during the hunting season, and the price of gasoline and its effect on hunter activity.

The papers in this special section are from a symposium on observational studies at the twelfth Wildlife Society conference in Madison, Wisconsin, USA. The purpose of that symposium was to share information on strategies for designing observational studies and analyzing data from them, and to foster a better understanding of the potential and the limitations of observational studies. We review observational studies, which are common in our science, and contrast them with experiments and, to a lesser extent, sample surveys. We identify features that distinguish each method of learning (borrowing freely from Johnson 2002, Johnson 2006, and other cited sources). We then indicate some complications that may arise when analyzing results of observational studies. We conclude by pointing out how the other articles in this special section can help scientists deal with observational studies.

CORNERSTONES OF MANIPULATIVE EXPERIMENTATION

Consider an example. Suppose one wants to determine how bobolinks (*Dolichonyx oryzivorus*) are affected by a treatment such as removal of woody vegetation in grasslands that had been invaded by trees and shrubs. The treatment effect (T) on a particular grassland can be defined as

$$T = Y_t(u) - Y_c(u), \tag{1}$$

where $Y_t(u)$ is the number of bobolinks in grassland u after the treatment, and $Y_c(u)$ is the number of bobolinks in that grassland if the treatment had not been applied. If the grassland is cleared, then one can observe $Y_t(u)$ but not $Y_c(u)$. If the treatment is not applied, then one can observe $Y_c(u)$ but not $Y_t(u)$. This leads to what has been termed the fundamental problem of causal inference: one cannot observe the values of $Y_t(u)$ and $Y_c(u)$ on the same unit (Rubin 1974, Holland 1986). That is, any particular grassland is either cleared or not.

Two solutions to this problem have been identified (Holland 1986). The first requires 2 units (u_1 and u_2, here grasslands) and the assumption that they are identical. Then the treatment effect T is estimated to be

$$T = Y_t(u_1) - Y_c(u_2), \tag{2}$$

where u_1 is treated and u_2 is not. This approach is based on the very strong assumption that the 2 grasslands, if not

[1] *E-mail: terry_shaffer@usgs.gov*

Table 1. Two examples of 4 grasslands, the value (no. of bobolink pairs) each would have if it were treated (i.e., woody vegetation removed), and the value it would have if it were not treated. In the example at left, all grasslands have identical values under each scenario. In the example at right, grasslands vary in values irrespective of treatment. In both examples, the effect of the treatment is 2 for all grasslands.

	Grasslands identical		Grasslands vary	
Grassland	**Value if treated**	**Value if not treated**	**Value if treated**	**Value if not treated**
1	3	1	3	1
2	3	1	5	3
3	3	1	2	0
4	3	1	6	4

Table 2. All possible samples of size 1 each, treated and untreated, from the population of grasslands that vary.

Treated grassland	Untreated grassland	Difference
1	2	0
1	3	3
1	4	−1
2	1	4
2	3	5
2	4	1
3	1	1
3	2	−1
3	4	−2
4	1	5
4	2	3
4	3	6

cleared, would have the same number of bobolinks, that is, $Y_c(u_2) = Y_c(u_1)$ and, if cleared, then $Y_t(u_2) = Y_t(u_1)$. We cannot test these assumptions, unfortunately, because one grassland had been cleared and the other had not. The assumption can be made more plausible by matching the 2 units as closely as possible or by believing that the units are identical. Physicists are more likely to believe that 2 molecules are identical than ecologists will think 2 grasslands the same, however.

The second solution has been termed statistical (Holland 1986). We can consider an expected, or average, causal effect T over all units in some population:

$$T = \mathrm{E}(Y_t - Y_c), \tag{3}$$

where, unlike with the first solution, different units can be observed. The statistical solution replaces the causal effect of the treatment on a specific unit, which is impossible to observe, by the average causal effect in the population, which is possible to estimate.

It is clear that a control, something to compare with the treated unit, is needed for either approach. In the statistical approach, randomization is often invoked. For example, if we are to compare bobolink numbers on a treated grassland and an untreated one, we could reach an erroneous conclusion if the grasslands differed greatly in size or vegetation structure. One way to protect against such possibly misleading outcomes is to decide at random which grassland is treated and which is not. Random assignment can be done in a controlled experiment but not in most observational studies.

Suppose in our example that there are 4 grasslands in our universe of grasslands of interest. And suppose, following the first solution, that they are identical: each would have 3 bobolinks if it were cleared and 1 bobolink if not (Table 1). Then, no matter which grassland we selected for treatment and which was the comparison, we would estimate the treatment effect to be 2, which is just right. But suppose that the grasslands themselves varied; we will maintain the nice simplifying assumption that the treatment effect would be 2 no matter what grassland we treated (Table 1). Then, if we treated one grassland and observed another as a control, there would be 12 possible combinations that could constitute our sample (Table 2). And our estimate of the treatment effect would vary depending on which grasslands we selected. For example, if we treated grassland 1 and grassland 3 served as a control, we would estimate our treatment effect as $3 - 0 = 3$. The 12 possible estimates of treatment effect range from −2 to 6. The average is 2, the correct value, but no possible sample would yield exactly that value.

So, even if we assigned treatments at random, it may just happen that one grassland would be large and have desirable vegetation structure (possibly grassland 4 in our example), and the other would be small with unsuitable vegetation (grassland 3). Such a sample would generate an estimated treatment effect (6) far from the correct value (2). This consideration leads us to the third important criterion for determining causation: replication. Repeating the randomization process and treatments on several grasslands makes it unlikely that grasslands in either group would be consistently more favorable to bobolinks. If we took a sample of size 2 for both treatment and control groups, there would be 6 possible samples, with estimated treatment effects ranging from −1 to 5 (Table 3). Note that samples with one grassland each in the treatment and in the control group yield estimates that vary around the true value (i.e., unbiased), but are very spread out (i.e., low precision; Fig. 1), whereas samples of size 2 in each group give estimates that cluster somewhat more closely around the true value (i.e., greater precision; Fig. 2). These then form the cornerstones for assessing the effect of some treatment with a manipulative experiment: controls, randomization, and replication (Fisher 1926).

One of the roles of randomization is to make variation

Table 3. All possible samples of size 2 each, treated and untreated, from the population of grasslands that vary.

Treated grasslands	Untreated grasslands	Difference
1,2	3,4	2
1,3	2,4	−1
1,4	2,3	3
2,3	1,4	1
2,4	1,3	5
3,4	1,2	2

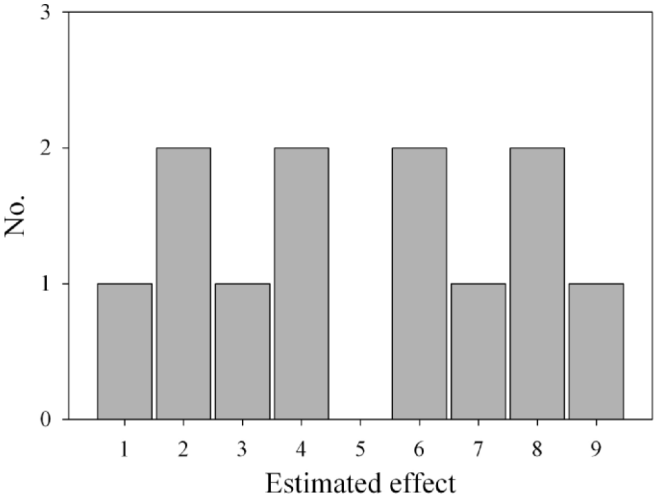

Figure 1. Distribution of estimated treatment effects from all possible samples of size 1 each, treated and untreated, from a simulated population of grasslands that vary.

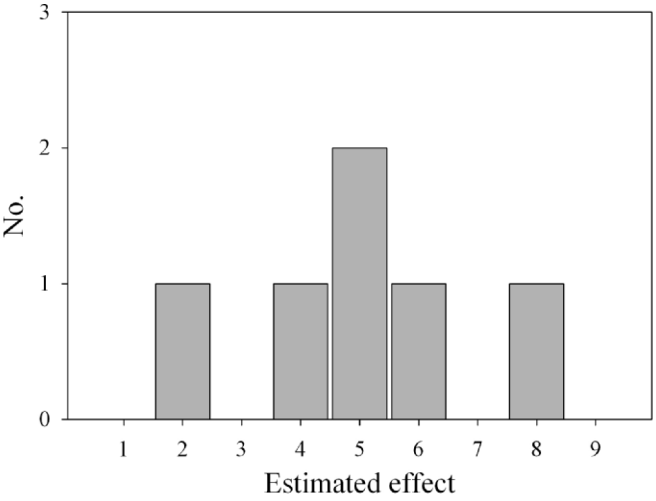

Figure 2. Distribution of estimated treatment effects from all possible samples of size 2 each, treated and untreated, from the simulated population of grasslands that vary.

among sample units, due to variables that are not accounted for, act randomly, rather than in some consistent and potentially misleading manner. Randomization thereby reduces the chance of confounding with other variables. Instead of controlling for effects of those unaccounted-for variables, randomization makes them tend to cancel one another, at least in large samples. In addition, randomization reduces any intentional or unintentional bias of the investigator. Because all outcomes are equally likely, randomization further provides an objective probability distribution for a test of significance (Barnard 1982).

Randomization by itself is not enough; replication is necessary for randomization to be useful. The desirable properties of randomization in the selection of study units are largely conceptual; that is, they pertain hypothetically to some long-term average. Randomization, for example, tends to make errors act randomly, rather than in some consistent fashion. However, in any single observation or study, the error may well be consistent. It is only through replication that large-sample or long-term properties hold. Replication provides 2 important benefits. First, it reduces error because an average of independent errors tends to be smaller than a single error. Replication serves to ensure against making a decision based on a single, possibly unusual, outcome of a treatment or measurement of a unit. Second, because we have several estimates of the same effect, we can estimate the error, as the variation in those estimates reflects error. We then can determine if the values of the treated units are unusually different from those of the untreated units. The validity of that estimate of error depends on the experimental units having been drawn randomly; thus, the validity is a joint property of randomization and replication.

EXPERIMENTS AND OBSERVATIONAL STUDIES CONTRASTED

Manipulative experimentation is an effective way to determine causal relationships. The investigator poses questions of nature via experiments, such as clearing woody vegetation. Because the investigator determines how the system is manipulated, the chance that something other than the treatment causes the observed results is reduced. Opportunities for complex interactions also are reduced because randomization lessens the chance that some unmeasured variable will affect the response in a way that is inconsistent among treatments.

Wildlife ecologists sometimes face severe difficulties meeting the needs of control, randomization, and replication in manipulative experiments. Many systems are too large and complex for ecologists to manipulate (Macnab 1983). Often putative treatments (sometimes referred to as exposure factors; e.g., oil spills) are applied by others, and wildlife ecologists are called in to evaluate their effects. In such situations, randomization is impossible and replication undesirable.

Observational studies lack one or more of the critical elements and, although they can be analyzed similarly to an experimental study (Cochran 1983), the investigator is less certain that the presumed treatment actually caused the observed response. In addition, observational studies are susceptible to situations involving complex interactions resulting from unmeasured variables that differentially affect the various levels of the treatment variable (see Riggs et al. 2008). Such interactions can greatly complicate the interpretation of treatment effects, but also can be informative because they reflect the inherent complexity of natural systems.

Sample surveys differ from experiments and, more subtly, from observational studies in that one endeavors either to estimate some characteristic over some domain, such as the number of mallards (*Anas platyrhynchos*) in the major breeding range in North America, or to compare variables among groups, such as the median age of hunters compared with nonhunters. In contrast, experiments and observational studies involve some sort of treatment.

ADJUSTING FOR INTRINSIC DIFFERENCES

Consider the difference between a single treated unit and a single untreated unit (grasslands, in our example): $Y_{ti} - Y_{cj}$, where unit *i* was treated and unit *j* was not. If the units were intrinsically identical (having value μ), then $Y_{ti} = \mu + T$, where *T* is the treatment effect, $Y_{cj} = \mu$, and so $Y_{ti} - Y_{cj} = T$. In our science, however, we cannot expect most units to be identical except for treatment effect, so we would have $Y_{ti} = \mu_i + T$ and $Y_{cj} = \mu_j$, where μ_i and μ_j are the intrinsic values associated with each unit. Then the difference $Y_{ti} - Y_{cj} = (\mu_i - \mu_j) + T$ is no longer a pure measure of *T* but also reflects the intrinsic difference between sample units.

If we cannot assume $(\mu_i - \mu_j) = 0$, there are 3 potential remedies. The first is to randomly select the units, so that the expected value of $(\mu_i - \mu_j)$ will be zero, and to replicate, so the actual value of the difference will be close to its expected value of zero. That is the approach taken in manipulative experiments. The second remedy involves selecting units that are as similar as possible, that is, $(\mu_i - \mu_j) \sim 0$. In an experimental context, this procedure is termed blocking, and it can markedly strengthen conclusions from an analysis. In observational studies, the scientists typically do not choose the units to receive treatment, but sometimes they can select units to compare to the treated units. Then selecting units as similar as possible to the treated units, except of course for the treatment, will tend to reduce the value of $(\mu_i - \mu_j)$. In observational studies, this is called matching, which is the counterpart of blocking in experimental studies. The third remedy is to attempt to estimate $(\mu_i - \mu_j)$ from available covariates associated with each unit (Eberhardt and Thomas 1991). In our grassland example, the number of bobolinks might vary in response to features such as field size, height and density of vegetation, and relative proportions of graminoid and forb vegetation. If that relationship can be modeled (e.g., with analysis of covariance [ANCOVA]; Milliken and Johnson 2002), then one could estimate μ_i and μ_j. Then *T* could be estimated by

$$\hat{T} = (Y_{ti} - Y_{cj}) - (\hat{\mu}_i - \hat{\mu}_j).$$

This is called statistical adjustment and is an example of model-based inference, in contrast to the design-based inference used in the first 2 approaches (Olsen et al. 1999). Matching requires no knowledge of the actual form of the relationship between the response variable (here, bobolink numbers) and the covariates; statistical adjustment does require that knowledge. Matching is sometimes hampered by an inability to find untreated units that are sufficiently similar to the treated ones. In contrast, statistical adjustment allows comparison units to be markedly different from treated units, as long as the functional dependence on covariates is well determined.

We briefly summarize these strategies as follows: blocking involves the prospective selection of (experimental) units and is used in experiments to increase precision of estimated treatment effects. Matching involves the retrospective selection of (comparison) units and serves the dual role of increasing precision and reducing bias in observational studies. Statistical adjustment involves the retrospective analysis of results.

Table 4. A population of 6 wetlands, the density of invertebrates in each, and its depth.

Wetland	Invertebrate density	Water depth
A	8	2
B	9	4
C	10	1
D	0	12
E	1	9
F	2	9

IMPLICATIONS FOR DATA ANALYSIS

The inability of investigators to randomly assign treatments in observational studies has important implications for data analysis. Techniques developed for analyzing data from experiments are commonly used with observational data as well. This practice seems perfectly fine on the surface but, as the following example shows, results can be seriously misleading.

Our example involves the introduction of fish into wetlands that support invertebrates used by foraging birds. The concern is that invertebrate density would be reduced as a result of predation by fish. We first considered how this question might be addressed with an experiment. Suppose 6 wetlands were available for study and that invertebrate density in the absence of fish ranged from 0 to 10 because of intrinsic differences among wetlands (Table 4).

Twenty outcomes, involving 3 treated wetlands and 3 control wetlands, would be possible from this experiment (Table 5). Random assignment of treatments would ensure that all outcomes were equally likely. If the treatment had no effect, the estimated treatment effect ($\hat{T}$) would range from −8 to 8, and the expected value, or average in this case, would be zero. The experiment would generate only 1 of the 20 possible outcomes, however, and the investigator would be faced with deciding whether that outcome was unusual enough to infer an effect of the treatment. That decision would be based on knowledge of the sampling distribution of $\hat{T}$ in the absence of a treatment effect (Fig. 3a); that knowledge would stem from the treatment being randomly assigned to wetlands. For example, a value of −8 for $\hat{T}$ would be improbable enough that, having observed that value, the investigator might justifiably conclude that the treatment was having an effect, even though it was not.

How does the above situation differ when treatment assignment is nonrandom, as is the case with observational studies? For illustration, we supposed that wetlands D, E, and F were 4 times more likely to receive fish than were wetlands A, B, and C. We represented this in terms of probabilities as P(D) = P(E) = P(F) = 4/15 and P(A) = P(B) = P(C) = 1/15, where P(·) was the probability that fish would be introduced into that wetland. Under this scenario, the same 20 outcomes would still be possible, but those outcomes would no longer be equally likely (Table 5). In

Table 5. All possible samples of size 3, to be treated, from the population of 6 wetlands (remaining 3 wetlands are controls), the probabilities of each sample under random and nonrandom assignment of treatments, and the estimated treatment effect ($\hat{T}$) for each sample.

	Assignment of treatment		
Treated sample	Random	Nonrandom	$\hat{T}$
A,B,C	0.05	0.002	8
A,B,D	0.05	0.012	1.3
A,B,E	0.05	0.012	2
A,B,F	0.05	0.012	2.7
A,C,D	0.05	0.012	2
A,C,E	0.05	0.012	2.7
A,C,F	0.05	0.012	3.3
A,D,E	0.05	0.062	−4.0
A,D,F	0.05	0.062	−3.3
A,E,F	0.05	0.062	−2.7
B,C,D	0.05	0.012	2.7
B,C,E	0.05	0.012	3.3
B,C,F	0.05	0.012	4
B,D,E	0.05	0.062	−3.3
B,D,F	0.05	0.062	−2.7
B,E,F	0.05	0.062	−2.0
C,D,E	0.05	0.062	−2.7
C,D,F	0.05	0.062	−2.0
C,E,F	0.05	0.062	−1.3
D,E,F	0.05	0.332	−8.0

fact, $\hat{T} = -8$ would be the most probable outcome when the treatment had no effect (Fig. 3b).

Of course, we would not know the true sampling distribution of $\hat{T}$ and would have to make some assumption about it to reach a conclusion about the effect of fish on invertebrate density. If we used analytical methods developed for experiments, our implicit assumption would be that the distribution matched the sampling distribution of the treatment effect (fish vs. no fish) under random assignment of treatments (Fig. 3a). Thus, if the study yielded $\hat{T} = -8$, we would conclude that the treatment had an effect, when in fact $\hat{T} = -8$ would be the most likely outcome in the absence of a treatment effect. Our above example demonstrates that the validity of using conventional methods developed for experiments (e.g., analysis of variance [ANOVA]) to analyze data from observational studies hinges on the assumption that treatments were randomly assigned.

The reader may be wondering how observational studies can lead to treatment assignments being as markedly nonrandom as our example would suggest. It is entirely plausible that, in the absence of random assignment of treatments, wetlands with fish would be deeper than wetlands without fish (Table 4). It is also reasonable to expect densities of invertebrates of some species to be greater in shallower wetlands. Thus, the interplay between wetland depth and both fish presence and invertebrate densities could give the illusion that invertebrate densities were being reduced by fish. Confounding like this commonly occurs in observational studies and points to the need for caution when interpreting results from observational studies. Matching might be useful in our example, except that we would need to include additional wetlands (i.e., shallower wetlands with fish and deeper wetlands without fish). Wetland depth is an obvious choice for the matching variable, but in many situations the choice of an appropriate matching variable, or variables, will be far from obvious. Statistical adjustment in the form of ANCOVA (Riggs et al. 2008) might also be effective in this situation, provided data were available for the confounding variables.

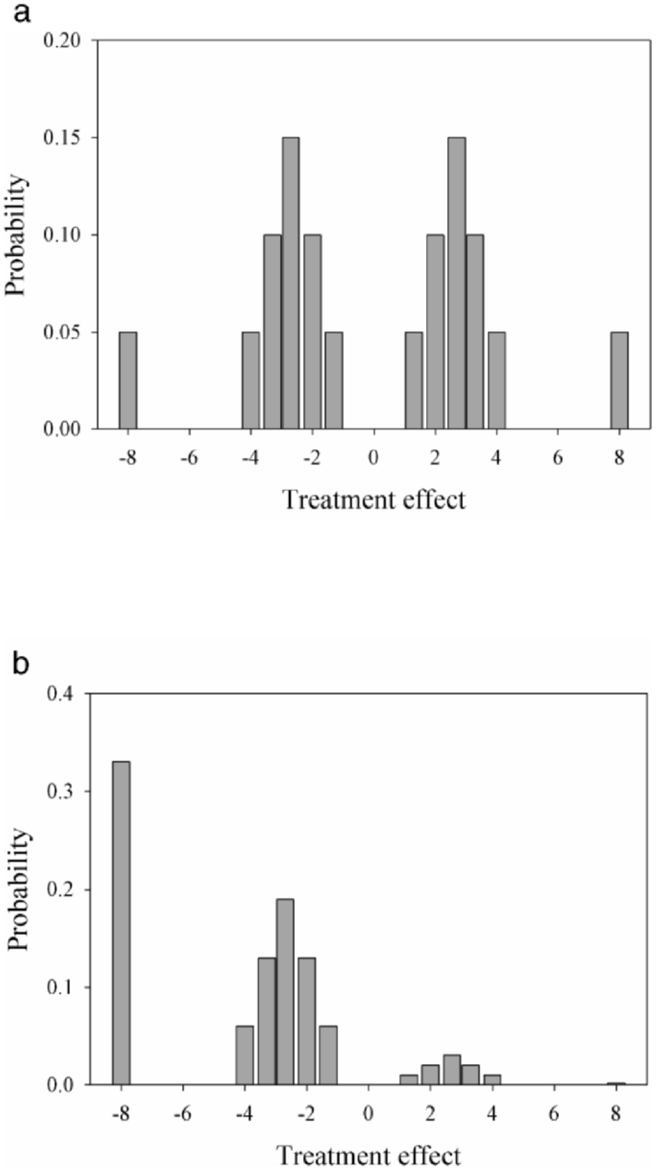

Figure 3. Sampling distribution of the treatment effect (fish vs. no fish) under (a) random and (b) nonrandom treatment assignment for samples of size 3 from the hypothetical population of wetlands, assuming that the treatment has no effect. Analysis methods developed for experiments are based on the distribution shown in (a), whereas observational studies may result in the distribution shown in (b).

ASSOCIATION VERSUS CAUSATION

Although observational studies and experiments differ in important ways, they share the common objective of elucidating causal relations (Cochran 1983). As we have seen, the importance of randomization in establishing causality cannot be overstated. Cochran (1983) recommended that results from observational studies be viewed with much skepticism, at least initially.

What steps can the investigator take to safeguard against reaching erroneous conclusions from observational studies? One important step is acknowledging that bias in observational studies is not only possible, but likely (Cochran 1983). Biases can be overt or hidden (Rosenbaum 2002). Overt biases are known to the investigator and can be neutralized through matching or analytic adjustment. Hidden biases are much more problematic because the information needed to detect or evaluate them is not available. Identifying and discussing all possible sources of bias and alternative explanations for the results observed can be helpful in understanding the potential degree of bias. As obvious as this advice may seem, it is often overlooked in practice (Cochran 1965).

Sensitivity analysis is a procedure for quantifying potential effects of hidden bias (Rosenbaum 2002). Sensitivity analysis poses the question: how much would inferences change in response to hidden biases of various magnitudes? Study conclusions are strengthened if results are found to be insensitive to biases of expected magnitude.

We illustrate the idea behind sensitivity analysis with the following example. Suppose that a waterfowl biologist has hypothesized that first-year breeding gadwall (*Anas strepera*) females are more likely to abandon their nests than are more experienced breeders, perhaps because first-year breeders are in poorer condition than older females. She investigates her hypothesis by locating nests, capturing and aging the females, and monitoring each nesting attempt to determine nest fate. Her analysis reveals that first-year breeders have an abandonment rate 3 times that of older females. One interpretation is that first-year breeders are in fact more prone to abandon their nests. An alternative explanation is that first-year breeders experience 3 times greater mortality while away from the nest. If off-nest mortality information is not available but the likelihood of a 3-fold difference in mortality rates is remote, then the original conclusion seems justified, at least in terms of that source of hidden bias. Sensitivity analysis attempts to address this type of issue in a formal way that recognizes sampling variability.

Another important consideration in analyzing and interpreting results of observational studies is the value of developing hypotheses and research questions a priori. Opportunities for data dredging (Burnham and Anderson 2002) are often great with observational studies because it may be relatively easy for the investigator to measure a large number of variables. Measuring numerous variables is the strategy in many observational studies, partly because the investigator may be unsure which variables will be useful for matching or as covariates during data analysis. Some observational studies are retrospective, meaning that the data were collected in the past (probably for some other purpose). Although such secondary data analyses can provide useful information, serious misuse occurs when the same data are used to identify a hypothesis and then test it (Williams et al. 2002). Pitfalls of data dredging are well established and have received much attention in the literature (e.g., Burnham and Anderson 2002).

In reference to what could be done to clarify the step between association and causation in observational studies, Sir Ronald Fisher said, "Make your theories elaborate" (Cochran 1965:252). We can gain insight into what Fisher meant by returning to our fish example. If invertebrate numbers were in fact reduced by fish, then not only would we expect higher densities in wetlands without fish, but we also would expect densities in wetlands with fish to be negatively related to fish abundance or perhaps to fish biomass. Taking this argument one step further, we also might expect effects of fish on invertebrate densities to be mitigated by the abundance of alternative food sources, such as small fry. Pursuing multiple lines of evidence and establishing consistency in results pertaining to various sub-hypotheses allows the investigator to weigh the evidence and systematically build a case for causation. Conversely, inconsistent results cast doubt over the nature of the relationship. In the end, professional judgment that considers all available evidence is necessary to render a decision regarding causality (McDonald et al. 2000).

Although it may seem obvious, the most important consideration, by far, with observational studies is to not regard the results of any one study as definitive. This is true for designed experiments, too, but is especially critical for observational studies. Meta-replication across space and time that involves multiple independent studies using a variety of techniques to address the same research question is paramount for establishing causal relationships with observational studies (Johnson 2002). Assessing the effect of smoking on human health provides a classic example of this. The fact that smoking causes lung cancer in humans was established through a series of observational studies that collectively provided overwhelming evidence, relative to the potential for bias (Rosenbaum 2002). In addition to smokers having greater incidence of lung cancer than nonsmokers, incidence was found to increase with exposure, strengthening the evidence that smoking causes cancer.

ANALYTIC COMPLICATIONS

Until now we have concerned ourselves with theoretical issues that make the analysis of data from observational studies problematic. With those issues in mind, we turn our attention to some practical matters that an investigator is likely to encounter when using techniques like ANOVA or ANCOVA to analyze data from observational studies. This is not an exhaustive list of complications, but it includes some that we have encountered or have witnessed going undetected or incorrectly treated in other studies. Our intent is not to provide detailed solutions to these problems but rather to alert the reader that the issues may be present. Most of these complications can be overcome, provided the investigator recognizes the problem. Assistance from a statistician with a background in linear models theory may be necessary to address some of the more difficult complications.

Unbalanced Data

In experimental design lingo, a design is said to be unbalanced if the number of experimental units varies among treatments or treatment combinations. A similar situation often occurs with observational studies. Unbalanced observational studies are not problematic so long as the investigator exercises caution when interpreting results. Consider an observational study involving 2 treatments (factors), say A and B. Suppose that both A and B occur at 2 levels (e.g., A might be sex of an individual study animal, and B might indicate whether the animal had been exposed or not exposed to some environmental influence), leading to 4 treatment combinations. We define the effect of each treatment as the difference between level 1 (treated) and level 2 (untreated). Denote the expected response (i.e., population mean) for a given treatment combination as μ_{ij}, where i denotes the level of A and j denotes the level of B. Let n_{ij} denote the number of animals observed and $\bar{y}_{ij}$ denote the mean response for a given treatment combination. The goal of the analysis is to draw inferences about the unobserved population means (μ_{ij}) based on values of the observed sample means ($\bar{y}_{ij}$).

When interactions are found to be unimportant, interest lies in estimating the main effect of each treatment, averaging across all levels of other treatments. For example, the average effect of treatment B is $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$, and an unbiased estimator of that effect is $(\bar{y}_{11} + \bar{y}_{21})/2 - (\bar{y}_{12} + \bar{y}_{22})/2$. Most analysis software will provide estimates of the main effects. For example, the LSMEANS statement of SAS (SAS release 8.2; SAS Institute, Cary, NC) will produce unbiased estimates of the marginal means. Confusion can arise with unbalanced data in that some software packages will also report $(n_{11}\bar{y}_{11} + n_{21}\bar{y}_{21})/(n_{11} + n_{21}) - (n_{12}\bar{y}_{12} + n_{22}\bar{y}_{22})/(n_{12} + n_{22})$. This statistic, which is a function of the sample sizes and is reported by the MEANS statement in several SAS procedures, may be useful in certain situations (e.g., when sample sizes are proportional to the number of animals in the various populations), but the statistic is a biased estimator of the main effect and generally is not useful. This confusion is easily avoided if analysts understand their software well. An analogous situation occurs with tests of hypotheses. Care is needed to ensure that the investigator understands what hypothesis is actually being tested.

Empty Cells

When one or more treatment combinations is not observed (i.e., $n_{ij} = 0$ for some i and j), the treatment structure is said to be incomplete. This situation (i.e., the empty cell problem [Hocking 1985]) is more problematic than unbalanced data because the investigator usually desires to make inferences about the unobserved and observed treatment combinations. For example, the quantity $([\mu_{11} + \mu_{21}]/2 - [\mu_{12} + \mu_{22}]/2)$ is still of interest, even if $n_{22} = 0$. However, if $n_{22} = 0$, this quantity cannot be estimated without making some assumptions. Similarly, the usual main effect hypotheses, such as H_0: $(\mu_{11} + \mu_{21})/2 = (\mu_{12} + \mu_{22})/2$, cannot be tested. Although it may not be obvious to the user, some software packages will produce statistics for testing a different hypothesis (e.g., $[\mu_{11} + \mu_{21}]/2 = \mu_{12}$ or $\mu_{11} = \mu_{12}$) when this situation arises. Investigators need to be alert for this possibility. Familiarity with one's data and knowledge of how one's software handles empty cells are extremely important.

Analysts should pay close attention to the degrees of freedom reported by their software for each term in an ANOVA or ANCOVA model. If the degrees of freedom for any term are fewer than expected, that is usually an indication of empty cells. For example, if treatment A has 4 levels and treatment B has 3 levels, then one should expect the $A \times B$ interaction term to have $(4 - 1)$ $(3 - 1) = 6$ degrees of freedom. If the software reports only 5 degrees of freedom, then some treatment combination is likely missing. Hopefully the software also will report that the usual main effects are not estimable. Depending on the number and arrangement of empty cells, refitting the model without the interaction term may allow us to estimate the main effects. However, we do not recommend this remedy unless one has firm evidence that the interaction between A and B is in fact negligible.

Covariate Affected by Treatment

Analysis of covariance provides a powerful tool for making model-based inferences from observational studies. A key assumption for ANCOVA modeling is that the covariate should not depend on the treatment (Milliken and Johnson 2002). It is not unusual with observational data, however, for the range in covariate values to vary with one or more of the treatments. When this happens, inferences concerning effects of the treatment can be perplexing and misleading, as the following example illustrates. In this example, we are interested in estimating the difference in accumulated body fat for some species under 2 sets of environmental conditions. Body fat measurements from a sample of 12 females and 15 males are available for analysis (Fig. 4). Males clearly have more body fat than females, although body fat increases with body size, making body size a logical covariate in the analysis. The relation between body fat and body size appears to be linear and exhibits the same rate of change for all 4 treatment combinations. Interactions between environmental condition and sex appear negligible. Thus, we can estimate the effect of environmental condition as the difference in estimated body fat between animals exposed to environment 1 and those exposed to environment 2 for either males or females. This estimate is typically computed using the mean value of the covariate (676 g). Results indicate that body fat is about 13 g greater for animals exposed to environment 1. Similarly, the effect of sex is the difference in body fat between males and females for either of the environments. The estimated sex effect, evaluated at the mean body size, turns out to be about −12, suggesting that males have less fat than females, which of course is not the case. Inspection of the data (Fig. 4) quickly

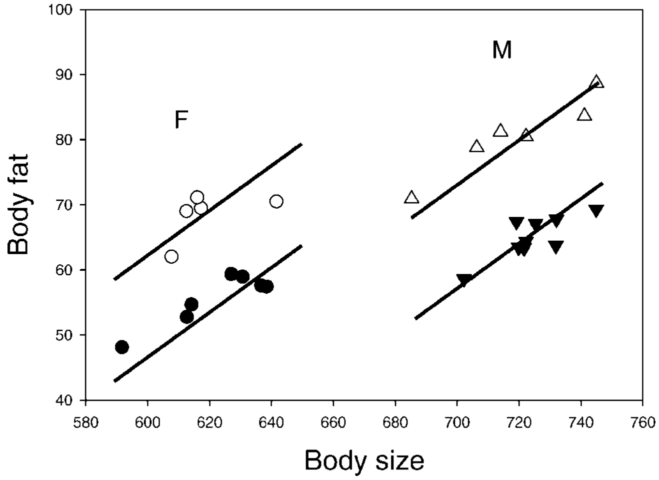

Figure 4. Hypothetical data showing body fat (g) of males and females exposed to 1 of 2 environments (environments distinguished by open vs. closed symbols) in relation to body size (g). Because body size varies with sex, standard covariance analysis gives the nonsensical result that males have less fat than females.

reveals that this nonsensical result is an artifact of the covariate varying by sex. The problem is not difficult to spot in a simple, contrived example like this, and where a simple graph can clarify the situation; however, it can easily go unnoticed when several covariates and several treatments are involved. Our example clearly illustrates the need to explore relationships among predictors before fitting models. Milliken and Johnson (2002) demonstrate a solution to the problem based on adjusting values of the covariate (Urquhart 1982).

Zero-Level Studies

The zero-level problem has a long history in experimental design. Zero-level problems also arise in observational studies. Suppose an investigator is comparing the effects of 2 grazing regimes, say season-long and rest–rotation regimes, on some wildlife response. Each regime is observed at 4 grazing intensities: high, medium, low, and none. At first glance, this appears to be a straightforward 2 × 4 treatment structure. The problem, however, is that because of the zero-intensity level there are only 7 treatment combinations. Standard ANOVA methods based on a 2 × 4 treatment structure are inappropriate because that approach uses a model that recognizes 8 treatment combinations, not 7. Zero-level problems can be accommodated by linear models theory (see Hocking 1985), but wildlife practitioners who have not previously encountered them may want to confer with a statistician.

Split-Plot Observational Studies

Split-plot experiments are characterized by 2 or more treatments that are assigned to 2 or more sizes of experimental units (Milliken and Johnson 1992). A classic example from agriculture involves the application of a fertilizer treatment to a plot of land, which is then subdivided into smaller plots to which different crop varieties are assigned. The entire setup is then replicated several times. The treatment assignment process results in more replicates for the variety treatment than for the fertilizer treatment. Researchers must take this distinction into account during the analysis, resulting in separate estimates for the split-plot and whole-plot error terms.

Split-plot situations are very common in observational studies, and they are often incorrectly analyzed. We return to the fish example to illustrate how a split-plot observational study can arise. Suppose that portions of individual wetlands are classified according to wetland zone, say deep marsh and shallow marsh. Wetland zone can then be viewed as a split-plot treatment that pertains to a portion of a wetland, whereas presence or absence of fish can be viewed as a whole-plot treatment that applies to the entire wetland. The same analytical considerations that apply to split-plot experiments (see Milliken and Johnson 1992) also apply to split-plot observational studies.

SPECIAL TOPICS AND CASE STUDIES

The 5 papers in this special section expand on and enhance some of the ideas we presented, and they illustrate effective methods for dealing with certain types of observational studies. Analysis of covariance models are perhaps the most widely used class of statistical models for analyzing observational data. In their simplest form, ANCOVA models combine the features of ANOVA and linear regression. The ability of ANCOVA models to simultaneously accommodate effects of multiple, continuous covariates and categorical covariates gives them broad appeal for use in observational studies. Riggs et al. (2008) examine this rich class of statistical models and expand on a number of issues that we considered briefly. Examples from 3 wildlife studies illustrate how ANCOVA models can be used to deal with effects of confounding variables and complex interactions. Riggs et al. also discuss model-fitting issues and available software for conducting analyses.

Many commonly used modeling techniques are predicated on the assumption of statistical independence of residuals. A frequent complication in the analysis of wildlife data is the occurrence of spatially or temporally correlated residuals. Small-scale spatial variation (i.e., spatially correlated residuals) may follow directly from a characteristic of the species (e.g., herding behavior) or because the animals respond to resources that vary spatially. Whatever the cause of the correlations, failure to account for them in the analysis can lead to improper inferences. Christman (2008) discusses the concept of spatial autocorrelation and provides an overview of 2 techniques (geostatistical modeling and lattice models) for incorporating small-scale variation in regression-type models. She identifies advantages and disadvantages of each approach and discusses implications of not modeling small-scale variation.

Wildlife ecologists often face situations in which many variables come into play, and sorting out the interactions among them all is a daunting task. Natural systems are complex and characterizing them requires multiple equations, in contrast to simpler processes that might be adequately described by a single equation. Grace (2008) provides an introduction to structural equation modeling, an analytic technique developed for summarizing relationships among multiple variables. He illustrates the methodology with an example involving a top predator, a mid-level predator, a prey species, and a habitat component. Grace also mentions available software and briefly describes some of the "opportunities for missteps" that should be avoided (Grace 2008:21).

Diseases among wildlife species are gaining increased attention from scientists and the public, not only because of the possible devastation they can cause to animal populations (e.g., chronic wasting disease among cervids, West Nile virus among certain birds) but also due to the potential for transmission to humans (e.g., avian influenza). Surveillance and monitoring of wildlife diseases therefore are critical activities. Traditional sample survey methods rarely are used, however, in part because the incidence of disease is low, at least in its initial stages. More often convenience sampling is used, such as monitoring animals shot by hunters. Nusser et al. (2008) address a key issue by simulating realistic scenarios to contrast the properties of real-world sampling schemes with more rigorous probability sampling.

Monitoring of natural resources such as animal populations represents a very common type of observational study. Recent years have seen the development of more sophisticated statistical treatments of data that result from monitoring activities. Link et al. (2008) tackle an intriguing problem: how to analyze data from 2 disparate programs that purport to monitor (approximately) the same thing. The North American Breeding Bird Survey monitors the size of bird populations in June, when most species are on territories and actively exhibiting breeding behaviors. The Christmas Bird Count is less rigorously systematic, but it provides numbers of birds observed during early winter. Link et al. (2008) indicate how the 2 surveys can be used in tandem, with a composite index from both of them providing more information than does either survey alone.

MANAGEMENT IMPLICATIONS

Scientists in the wildlife profession, as in many other professions, are envious of the physical sciences, in which manipulative experimentation plays a strong role in rapidly advancing understanding. Numerous manipulations are done in our profession, but they are done by managers, and the manipulations often lack controls, randomization, or replication. We rarely take full advantage of the lessons we could learn from those manipulations (Macnab 1983). Far too many manipulations of habitat, for example, are conducted without sufficient follow-up monitoring of treated and control units to evaluate the actual effects of the manipulation. Instead, often it is simply assumed that the consequences will be what they were expected to be. We should formally adopt the simple yet key ideas of adaptive resource management and make management decisions based on our understanding, monitor the consequences of the management actions, and revise our understanding based on those results (Walters 1986, Williams et al. 2002). As Macnab (1983) noted, assumptions underlying a management action should be treated as hypotheses, rather than facts. Our ability to learn from management actions can be enhanced by striving for evaluations that whenever possible meet the 3 cornerstones of experimentation. But often wildlife professionals must settle for observational studies on systems that are affected by numerous influences, only some of which are known and fewer are measured.

We indeed have a more difficult science to understand than physics. Rather than feel self-pity about our inability to manipulate the systems we investigate, we should capitalize on the realism that observational studies provide and use the best methods available to learn about the systems. In our field, carefully controlled experiments rarely capture the full range of variation that occurs in nature, so observational studies clearly offer more realistic settings than do experiments. With increased realism, however, comes added complexity and the danger of misleading results. We hope the articles in this special section will enhance the ability of wildlife and other natural resource professionals to more fully exploit the learning opportunities provided by observational studies.

ACKNOWLEDGMENTS

We thank the authors who contributed papers to this special section. The Biometrics Working Group of The Wildlife Society sponsored the symposium on observational studies, and members provided helpful advice and encouragement. We are grateful to D. A. Buhl, R. R. Cox, Jr., J. R. Fieberg, R. A. Gleason, J. B. Grace, W. A. Link, and M. R. Riggs for comments that improved the manuscript, and we are grateful to B. R. Euliss for preparing the figures.

LITERATURE CITED

Barnard, G. A. 1982. Causation. Pages 387–389 *in* S. Kotz and N. L. Johnson, editors. Encyclopedia of statistical sciences. Volume 1. Wiley, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Christman, M. C. 2008. Statistical modeling of observational data with spatial dependencies. Journal of Wildlife Management 72:23–33.

Cochran, W. G. 1965. The planning of observational studies of human populations (with discussion). Journal of the Royal Statistical Society, Series A 128:234–265.

Cochran, W. G. 1983. Planning and analysis of observational studies. Wiley, New York, New York, USA.

Eberhardt, L. L., and J. M. Thomas. 1991. Designing environmental field studies. Ecological Monographs 61:53–73.

Fisher, R. A. 1926. The arrangement of field experiments. Journal of Ministry of Agriculture of Great Britain 33:503–513.

Grace, J. B. 2008. Structural equation modeling for observational studies. Journal of Wildlife Management 72:14–22.

Hocking, R. R. 1985. The analysis of linear models. Wadsworth, Belmont, California, USA.

Holland, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association 81:945–960.

Johnson, D. H. 2002. The importance of replication in wildlife research. Journal of Wildlife Management 66:919–932.

Johnson, D. H. 2006. The many faces of replication. Crop Science 46: 2486–2491.

Link, W. A., J. R. Sauer, and D. K. Niven. 2008. Combining Breeding Bird Survey and Christmas Bird Count data to evaluate seasonal components of population change in northern bobwhite. Journal of Wildlife Management 72:44–51.

Macnab, J. 1983. Wildlife management as scientific experimentation. Wildlife Society Bulletin 11:397–401.

McDonald, T. L., W. P. Erickson, and L. L. McDonald. 2000. Analysis of count data from before–after control–impact studies. Journal of Agricultural, Biological, and Environmental Statistics 5:262–279.

Milliken, G. A., and D. E. Johnson. 1992. Analysis of messy data. Volume I: designed experiments. Chapman and Hall, London, United Kingdom.

Milliken, G. A., and D. E. Johnson. 2002. Analysis of messy data. Volume III: analysis of covariance. Chapman and Hall, London, United Kingdom.

Nusser, S. M., W. R. Clark, D. L. Otis, and L. Huang. 2008. Sampling considerations for disease surveillance in wildlife populations. Journal of Wildlife Management 72:52–60.

Olsen, A. R., J. Sedransk, D. Edwards, C. A. Gotway, W. Liggett, S. L. Rathbun, K. H. Reckhow, and L. J. Young. 1999. Statistical issues for monitoring ecological and natural resources in the United States. Environmental Monitoring and Assessment 54:1–45.

Pearl, J. 2000. Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, United Kingdom.

Riggs, M. R., K. J. Haroldson, and M. A. Hanson. 2008. Analysis of covariance models for data from observational field studies. Journal of Wildlife Management 72:34–43.

Rosenbaum, P. R. 2002. Observational studies. Second edition. Springer, New York, New York, USA.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66:688–701.

Urquhart, N. S. 1982. Adjustment in covariance when one factor affects the covariate. Biometrics 38:651–660.

Walters, C. 1986. Adaptive management of renewable resources. Macmillan, New York, New York, USA.

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic Press, New York, New York, USA.

Associate Editors: Shaffer and Johnson.